A computational approach to lexical polysemy in Ancient Greek

McGillivray, Barbara

2019-08

McGillivray , B , Hengchen , S , Lähteenoja , V E , Palma , M & Vatri , A 2019 , ' A computational approach to lexical polysemy in Ancient Greek ' , Digital scholarship in the þÿ h u m a n i t i e s , v o l . 3 4 , n o . 4 , p p . 8 9 3 9 0 7 . h t t p s : / / d o i . o r g / 1 0 . 1 0 9 3 / l l c /

http://hdl.handle.net/10138/333162
https://doi.org/10.1093/llc/fqz036

acceptedVersion

A computational approach to lexical polysemy in Ancient Greek

Barbara McGillivray
The Alan Turing Institute, United Kingdom
Theoretical amd Applied Linguistics, Faculty of Modern and Medieval Languages, University of Cambridge, United Kingdom

Simon Hengchen
COMHIS, Department of Digital Humanities, University of Helsinki, Finland

Viivi Lähteenoja
Doctoral Programme for Philosophy, Arts, and Society, University of Helsinki, Finland

Marco Palma
Department of Statistics, University of Warwick, United Kingdom

Alessandro Vatri
Wolfson College, University of Oxford, United Kingdom

1. Introduction: lexical polysemy and semantic change

Language is a complex and dynamic system. If we consider word meaning, which is the scope of lexical semantics, we observe that some words have several meanings, thus displaying *lexical polysemy*. For example, *head* in English means the body part and a group leader. This situation may change over time, giving rise to *semantic change*, whereby a word acquires new meanings or loses existing ones. For example, *tweet* refers to the chirp of small birds, but in recent years it has been used also to refer to a post on Twitter (cf. e.g. entry 'tweet, v.' in the Oxford English Dictionary). Sometimes two meanings coexist before one takes over, as in the case of the French word *témoin* 'person giving testimony', which originated from the Old French *tesmoin* 'testimony/person giving testimony', and derived from the Classical Latin word *testimonium* 'testimony' (Koch 2016: 25).

Semantic change phenomena are grounded in universal cognitive processes combined with cultural, linguistic, and historical factors (cf. Koptjevskaja-Tamm 2016). Its mechanisms have been studied by linguistics fields for some time, starting with diachronic linguistics in the 19th century and continuing with semiotics and cognitive semantics in the 20th century. Beyond linguistics research, understanding the way in which words change their meaning benefits a range of humanistic fields. As a consequence of semantic change, words used to refer to certain concepts can be different over time. The English adjective *gay* had the original meaning of 'happy' and 'joyful', but over the past century it has increasingly been used to refer to homosexual men (see the entry 'gay, adj., adv., and n.' in the

Fig. 6: Output of SCAN for the word *parabolê*.

6. Discussion and Conclusion

The ability to use computers to scale up manual analyses of word semantics over time is going to open up exciting opportunities in humanistic research. In this article we report on research that takes the first steps in this direction for Ancient Greek. This offers us the opportunity to reflect on the challenges and the future research directions in this area, as well as more generally on the value of computational modelling in Humanities research.

We know that computational modelling of polysemy and semantic change in natural languages is a difficult task, as proven by the fact that this is still an open problem in computational linguistics. For historical and Ancient languages specifically, the task is made even harder by a combination of factors, as shown by the analyses we reported on in this article.

When SCAN (Frermann and Lapata 2016), a statistical model of semantic change developed for relatively recent stages of a modern language (English), is applied to the Ancient Greek corpus, several adjustments had to be made. The features of the corpus meant that genre balance could not assumed, leading us to experiment with the GASC model (Perrone *et al.* 2019) that considers the texts' genre characteristics and their effect on polysemy. As McGillivray (2014) noted for computational processing of Latin data, another question concerned data sparsity, , although the Diorisis corpus is one of the largest annotated corpora of Ancient Greek available. This was addressed by a periodization of the corpus by century, so that the model was provided with enough data to draw its inferences from. Despite these adjustments, a close study of *kosmos* shows that data sparsity is still a challenge, as shown by the fact that the target words appear several times in their contextual list of words.

Further linguistic considerations are also in order. Semantic polysemy is a complex phenomenon, which encompasses cultural and historical events, literary and stylistic conventions, but also different linguistic patterns. For example, one meaning can emerge as an extension of another, leading to fuzzy boundaries. The example of *paradeisos* illustrates this, as the religious sense is closely related (also distributionally) to the original meaning of 'garden'. This has important implications for computational implementations, which can partially capture even such subtle changes. On the other hand, as for *parabolê*, when the emerging meaning is not distinguishable from the original one based on the context of the word, but based on other factors, the task becomes harder for distributional systems.

Finally, from a methodological viewpoint, the question of how to evaluate computational models of semantics is still open. We have shown the challenges of comparing computational output with expert annotation, and of setting correct expectations regarding the acceptable level of misalignment between expert analysis and automatic analysis. In spite of these limits, we have shown the potential of such statistical models to allow researchers to quantify complex semantic phenomena, thus going beyond the traditional account of historical semantics in terms of examples of first attestations of meanings (as reported in historical dictionaries, for instance). Such quantitative accounts make it possible, among other things, to gain insights into the dynamics between different meanings of a word, and ultimately, to place this in its wider historical and causal context.

Acknowledgements

This work was supported by The Alan Turing Institute [EPSRC grant EP/N510129/1, SF0042 to BMcG].

Contributions

BMcG designed and oversaw the study, designed the genre analysis described and the evaluation approach, contributed to the design of the computational model, drafted sections 1, 4, 5, and 6, reviewed all other sections, and acquired the funding; SH drafted sections 1, 2 and 6, contributed to the design of the computational model, and reviewed all other sections. VL annotated the sample for *kosmos* and interpreted its results, and drafted section 5. MP contributed to the design and implementation of the computational model, drafted sections 4.2 and 5, and prepared figures 2, 3, 4, 5, and 6. AV collected and processed the corpus data, annotated the sample for *mus* and *harmonia* and interpreted their results, contributed to the design of the computational model, conducted the genre analysis, and drafted sections 3 and 5. All authors gave final approval for publication.

References

Allan, D.J. (1980). Ἀναγιγνώσκω and Some Cognate Words. The Classical Quarterly n. s. 30: pp. 244–51.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan), pp. 993-1022.

Boschetti, F. (2010). A corpus-based approach to philological issues. Ph.D. thesis, University of Trento.

Christidis, A.-P. (ed.) (2007). A history of Ancient Greek: from the beginnings to Late Antiquity. Cambridge University Press, Cambridge.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In Resnick, L. B.; Levine, J. M., (eds.), Perspectives on socially shared cognition, American Psychological Association.

Clarke, M. (2010). Semantics and Vocabulary. In Bakker, E. J. (ed.). A Companion to the Ancient Greek Language, pp. 120–33. Wiley-Blackwell, London.

Dubossarsky, H., Tsvetkov, Y., Dyer, C. and Grossman, E. (2015). A bottom up approach to category mapping and meaning change. In NetWordS (pp. 66-70)

Dubossarsky, H., Weinshall, D. and Grossman, E. (2017). Outta control: Laws of semantic change and inherent biases in word representation models. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing pp. 1136-1145.

Foxlee, N. (2015). From Analogue to Digital: Conventional and Computational Approaches to Studying Conceptual Change. Retrieved online at https://www.academia.edu/20046437/From_Analogue_to_Digital_Conventional_and_Computational_Approaches_to_Studying_Conceptual_Change_2015_ on September 12th, 2018.

Frermann, L. and Lapata, M. (2016). A Bayesian model of diachronic meaning change. Transactions of the Association for Computational Linguistics, *4*, pp.31-45

Hamilton, W.L., Leskovec, J. and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.0909

Harris, Z.S. (1954). Distributional structure. Word, 10(2-3), pp.146-162

Hengchen, S. (2017). When does it mean? Detecting semantic change in historical texts. Ph.D. thesis, Université libre de Bruxelles.

Jenset, G. B. and McGillivray, B. (2017). Quantitative Historical Linguistics. A corpus framework. Oxford University Press, Oxford.

Koch, Peter (2016). Meaning change and semantic shifts. In Juvonen, Päivi & Koptjevskaja-Tamm, Maria (eds). The Lexical Typology of Semantic Shifts. De Gruyter, Berlin.

Koptjevskaja-Tamm, Maria (2016). "The lexical typology of semantic shifts. An introduction". In Juvonen, Päivi & Koptjevskaja-Tamm, Maria (eds). The Lexical Typology of Semantic Shifts. De Gruyter, Berlin.

Koselleck, R. (2006). "Crisis". Trans. Michaela W. Richter. Journal of the History of Ideas, 67(2): 357-400.

Kutuzov, A., Øvrelid, L., Szymanski, T., & Velldal, E. (2018). Diachronic word embeddings and semantic shifts: a survey. Arxiv: http://arxiv.org/abs/1806.03537

Kyrtatas, D. J. (2007). Παράδεισος. In Christidis, A.-P. (ed.). A history of ancient Greek: from the beginnings to Late Antiquity, 1137–40. Cambridge University Press, Cambridge.

Leiwo, M. (2012). Introduction: Variation with Multiple Faces. In Leiwo, M., Halla-aho, H., and Vierros, M. (eds.). Variation and change in Greek and Latin, 1–11. Suomen Ateenan-instituutin säätiö, Helsinki.

Liddell, H.G. and Scott, R. (1996). A Greek English Lexicon. 9th edition with revised supplement. Oxford University Press, Oxford.

McGillivray, B. (2014). Methods in Latin Computational Linguistics. Brill, Leiden.

Nerlich, B. (1992). Semantic Theories in Europe, 1830-1930: From etymology to contextuality, volume 59. John Benjamins Publishing, Amsterdam.

Perrone, V., Palma, M., Hengchen, S., Vatri, A., Smith, J.Q., McGillivray, B. (2019). GASC: Genre-Aware Semantic Change for Ancient Greek. Arxiv: https://arxiv.org/abs/1903.05587

Pollitt, J.J. (1974). Ancient View of Greek Art (pp. 247-54). New Haven: Yale University Press.

Reisig, K. (1839). Vorlesungen über lateinische Sprachwissenschaft. Lehnhold, Leipzig.

Rodda, M.A., Senaldi, M.S. and Lenci, A. (2017). Panta Rei: Tracking Semantic Change with Distributional Semantics in Ancient Greek. In CLiC-it/EVALITA.

Setatos, M. (2007). Semantic Change. In Christidis, A.-F. (ed.). A History of Ancient Greek: from the Beginnings to Late Antiquity, 667–76. Cambridge University Press, Cambridge.

Schmid, H. (1995a). Probabilistic Part-of-Speech Tagging Using Decision Trees. Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

Schmid, H. (1995b). Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin.

Tahmasebi, N., Borin, L., Jatowt, A. (2018). Survey of Computational Approaches to Lexical Semantic Change. arXiv preprint arXiv:1811.06278.

Trier, J. (1931). Der Deutsche Wortschatz im Sinnbezirk des Verstandes: von den Anfängen bis zum Beginn des 13. Jahrhunderts (Vol. 31). Winter.

van Groningen, B.A. (1963). Ἔκδοσις. Mnemosyne 16: 1–17.

Vatri, A. (2016). Between Song and Prose: The Meaning(s) of *Harmonia* in Aristotle's *Rhetoric* and *Poetics*. Rhetorica 34: 372–92.

Vatri, A. and McGillivray, B. (2018). The Diorisis Ancient Greek Corpus. Research Data Journal for the Humanities and Social Sciences. https://brill.com/view/journals/rdj/aop/article-10.

1163-24523666-01000013.xml

Vatri, A., Lähteenoja, V. and McGillivray, B. (2019). Ancient Greek Semantic Change - Annotated Datasets and Code. figshare. doi:10.6084/m9.figshare.c.4445420.